Week 8

~~Hour 1: More on polynomial fits. The AIC~~

~~Hour 2: Dummy Variables~~

Hour 3: Interactions

Interactions.

So far we have extended simple regression in the following ways:

- Multiple regression: More than 1 explanatory variable.
- Polynomial terms: Non-linear functions of a variable.
- Dummy variables: Categorical variables.

Now here's one more,

- Interactions: Combining variables.

Interactions are a way of combining explanatory variables to make new regression terms.

An interaction between two explanatory variables, (let's call them X_1 and X_2 is useful whenever the effects of X_1 and X_2 are not ***additive***.

What does additive mean?

Recall from last week when we introduced multiple regression we treated/assumed effects were additive.

If the response Y increases with x_1 at a (marginal) rate β_{x1} and Y increases with x2 at rate β_{x2} then increasing both x_1 and x_2 by 1 should increase Y by $\beta_{x1} + \beta_{x2}$.

Example:

Let the yield for a certain crop respond to average daily rain (rain) and average daily high temperature (temp).

The yield increases by 3 units for each 1 C increase in Temp,

and increases by 5 units for each 1 mm increase in Rain.

In math terms,

$\beta_{x1} = 3, \beta_{x2} = 5$

But what if that's not how this crop works?

What if it needs BOTH heat and rain TOGETHER to thrive?

Then an increase in heat, holding rain constant, should produce only a small effect.

Likewise, an increase in rain, holding temperature constant, would only increase yield a little.

... But increasing BOTH Temp and Rain would produce an effect that ***larger than the sum*** of either effect alone.

This extra effect from the combination of Temp and Rain together is called an ***interaction effect***.

We can capture the interaction effect with a regression equation like this:

$$\text{Yield} = \beta_0 + \beta_1(\text{Rain}) + \beta_2(\text{Temp}) + \beta_{12}(\text{Rain X Temp})$$

...where β_{12} is the parameter for the interaction effect.

It's positive because the effect from rain and temperature together is more than the individual effects.

If there is an interaction, the effects are not additive.

With a dataset, such a model with an interaction could look like this:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.5558    67.5770  -0.067    0.946
rain          0.4462     3.0610   0.146    0.884
temp          1.3371     2.6653   0.502    0.616
rain:temp     0.8867     0.1192   7.441 7.96e-13 ***
```

The easiest way to determine if an interaction term is appropriate is a hypothesis test.

Include the interaction in question in your model, and check if the associated p-value is small, or if the AIC is smaller with the model.

Like polynomial terms, interactions introduces co-linearity.

Including the interaction between two variables A,B, means that both variables appear in two terms each.

So both the ***main effect*** terms as well as the interaction term will have high VIF.

```
> vif(mod1)
    rain     temp
1.008627 1.008627
> vif(mod2)
     rain      temp rain:temp
 21.48286   7.69843  30.32871
```

What this means is that it's sometimes hard to isolate the effects on the response from each term.

However, as usual, a high VIF does not impact prediction strength.

One other thing:

It's a good practice to include the ***main effects*** for any interaction terms being included.

This is true even if one or both of the main effects is redundant.

For example, if the model

$$\text{Yield} = \beta_0 + \beta_1 \text{Rain} + \beta_{12}(\text{Rain} * \text{Temp})$$

had a better AIC than the more complex model

$$\text{Yield} = \beta_0 + \beta_1(\text{Rain}) + \beta_2(\text{Temp}) + \beta_{12}(\text{Rain} * \text{Temp})$$

…then it would still be preferable to use the model that includes both the Rain main effect and the Temp main effect.

This is called the ***Principle of Hierarchy*** – a model that includes an interaction of two (or more) variables should also include those variables on their own.

The principle of hierarchy is mostly for the sake of being able to explain the model.

Things get ugly when you mess with hierarchy.

Interactions with categorical variables

Sometimes a categorical doesn't just change a response in a vacuum, sometimes its effect on the response depends on other variables.

Consider the dataset Dragons.csv , in which we are trying to model the weight of bearded dragons as a function of age, length, and sex. Sex is a categorical variable, so it gets given a dummy variable in the regression.

Our regression equation is:

(Weight) =

-551.1 + 17.1(Age) +34.3(Length) + 4.9(Female)

Meaning that a dragon with 0 age, 0 length, and not female has an average weight of -551.1.

Every year of Age adds 17.1 grams of weight,

Every cm of length adds 34.3, and being female adds 4.9 to the weight of a typical beardie.

There are some biology based objections to this model:

1. Other proportions increase too, so weight should increase by length cubed if anything.

2. Female bearded dragons increase in weight faster than males do. It's not like they're almost 4.9 grams heavier.

Objection 1 can be solved with a ***polynomial fit***, where we include both a length and a length-cubed term in the model.

Objection 2 is essentially saying there is ***an interaction*** between sex and length.

An interaction between a dummy variable (sex = M) and a continuous one (length) is possible.

The parameter for such an interaction is interpreted as a slope (over the continuous), but specific to that group.

Consider the more complicated model for beardie weight:

$$\text{Weight} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Length}) + \beta_3(\text{Length}^3) + \beta_4(\text{Sex} = \text{M}) + \beta_{14}(\text{Age} * (\text{Sex}=\text{M}))$$

For females,

the intercept is $\boldsymbol{\beta_0}$ **,** and weight increases by $\boldsymbol{\beta_1}$ per year.

For males,

the intercept is $\boldsymbol{\beta_0 + \beta_4}$, weight increases by $\boldsymbol{\beta_1 + \beta_{14}}$ per year.

Flipping this around,

β_1 Is the added weight at age 0 for female dragons, and

β_{14} Is the added weight per year for male dragons.

If the weight gain per year is the same for each sex, β_{14} will be zero. We can test that by looking at a regression summary.

```
> mod = lm(weight ~ size + size^3 + age + sex + age:sex
> summary(mod)
                                                , data=dragons)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -769.064     68.017 -11.307  2.2e-15 ***
size          70.908      2.408  29.442  < 2e-16 ***
age          116.656      9.498  12.282  < 2e-16 ***
sexM          50.606     66.136   0.765 0.447766
age:sexM     -55.999     14.873  -3.765 0.000439 ***
```

A few more comments on categorical variables and interactions:

If we were to look at an interaction involving a categorical variable with more than two groups, each dummy variable would have its own interaction term in the regression.

The would be interpreted as the increase/decrease in slope for each group compared to the slope of the baseline group.

You can also make an interaction term between two categorical variables.

In this case, the dummies from one categorical variable get an interaction term with the dummies from the other.

Example: An interaction between a 6 group categorical and a 4 group categorical variable would introduce

(6 - 1) X (4 – 1) = 15 interaction terms,

And would cost 15 DF to include.

A few more comments on interactions in general:

You can have interactions between three variables as well. These are called third-order interactions, and they generally aren't worth the effort.

Another option is to try an interaction between a polynomial term and a linear term in a regression.

Models can come from anywhere.

Stepwise

Now that we've introduced interactions, there are so many options for building statistical models that we need a method to work through many possibilities quickly.

The stepwise method is one such method.

For the stepwise method, you need..

- One response variable

- A list of explanatory variables that could included in the model.

- A criterion for evaluating the quality of a model.

Ideally, stepwise will find the combination of those explanatory variables that produces a model that gets the best score for the selected criterion.

The most popular R function for stepwise is stepAIC() in the MASS package.

The default criterion used is ***AIC***, but it's easy to change it to BIC or R-squared*.

stepAIC inputs a 'starting model', which includes all the terms you want to consider. It outputs a 'final model' which you can use just like anything you would get from lm()

*Don't use R-squared. Seriously. Just don't.

Consider the gapminder dataset, and our model of birth rates from before.

When we found the best model using VIF, and again with AIC, we didn't consider interactions.

For the six variables (all continuous) we were considering before, there are 5+4+3+2+1=15 possible interaction terms to consider.

We won't consider any polynomial terms.

For each term, we can either include it or not, independently of the other terms included.

That means there are 2^{21}, or ~2 million possible models we can build using the 6 main effects and 15 interactions.

Rather than trying to manually find the best combination, we can feed this information into a stepwise function in R and it will find one for us.

Let's try it:

Input

```
> library(MASS)
> gapminder = read.csv("gapminder.csv")
>
> gapminder = subset(gapminder, !is.na(GINI) & !is.na(HDI) &
+ !is.na(agri_in_gdp) & !is.na(GDPpercap) & !is.na(health_spending))
>
> mod_start = lm(birth_rate ~
+ (GINI + HDI + female_work + agri_in_gdp + GDPpercap + health_spending)^2,
+ data=gapminder)


> mod_final = stepAIC(mod_start)
> summary(mod_final)
```

Summary of output:

Main Effects

Interactions

```
Coefficients:
                                  Estimate        Pr(>|t|)
(Intercept)                     -3.331e+01        0.423492
GINI                             1.451e+00        0.000192 ***
HDI                             -6.416e+01        0.166437
female_work                      5.057e-01        0.367017
agri_in_gdp                      5.677e+00        0.001023 **
GDPpercap                        1.508e-04        0.885961
health_spending                  4.336e-02        0.008339 **
GINI:female_work                -1.503e-02        0.003215 **
GINI:agri_in_gdp                -2.806e-02        0.016858 *
HDI:female_work                  1.189e+00        0.076021 .
HDI:agri_in_gdp                 -3.785e+00        0.007802 **
female_work:agri_in_gdp         -3.333e-02        0.012153 *
female_work:health_spending     -1.174e-03        9.47e-05 ***
agri_in_gdp:health_spending     -1.520e-03        0.179806
GDPpercap:health_spending        3.395e-06        0.031976 *
```

14 of 21 possible terms have been included.

A 14-term regression may work for some situations, but we may also want a simpler model. That is, one with fewer terms.

To do this, we need to use a criterion with a larger penalty for complexity, such as the ***Bayesian Information*** Criterion (BIC).

To use this we change the 'k' setting in stepAIC, which is penalty per term. The default value for 'k' is 2.

```
> mod_final_bic = stepAIC(mod_start, k = log(150))
> summary(mod_final_bic)
```

BIC-based output:

Main Effects

Interactions

Removed

```
Coefficients:
                                 Estimate      Pr(>|t|)
(Intercept)                    -4.197e+01      0.320910
GINI                            1.504e+00      0.000137 ***
HDI                            -4.927e+01      0.279462
female_work                     4.933e-01      0.389337
agri_in_gdp                     6.058e+00      0.000546 ***
GDPpercap                      -9.356e-04      0.198188
health_spending                 3.290e-02      0.019395 *
GINI:female_work               -1.529e-02      0.003171 **
GINI:agri_in_gdp               -2.923e-02      0.014609 *
HDI:female_work                 1.200e+00      0.079360 .
HDI:agri_in_gdp                -4.684e+00      0.000486 ***
female_work:agri_in_gdp        -3.201e-02      0.016818 *
female_work:health_spending    -1.166e-03      0.000115 ***
GDPpercap:health_spending       4.558e-06      0.001489 **
```

```
agri_in_gdp:health_spending
```

13 of 21 possible terms have been included.

Three big drawbacks to the stepwise method:

1. It can only consider terms that you specify. It won't try things like additional polynomial terms, interactions, or transformations for you.

2. It doesn't actually try every possible candidate model, so there is a chance that a better model exists that the stepwise method will miss.

3. It blindly applies the given criterion without regards to other concerns like non-linear fits, and influential outliers, co-linearity, and hierarchy.

In short, the stepwise method is…

not a replacement for human judgement.